DOCUMENT RESUME

ED 109 229 TM 004 690

AUTHOR Steinfatt, Thomas M.
TITLE A Comment on the "Known-Interval" Scale as a Measure of Attitude.
PUB DATE Dec 74
NOTE 18p.; Paper presented at the Annual Meeting of the Speech Communication Association (Chicago, Illinois, December 27-30, 1974). Not available in hard copy due to marginal legibility of original document

EDRS PRICE MF-$0.76 PLUS POSTAGE. HC Not Available from EDRS.
DESCRIPTORS *Attitude Tests; Predictive Validity; Statistical Bias; *Testing Problems; *Test Reliability; *Tests of Significance; *Test Validity
IDENTIFIERS *Known Interval Scale

ABSTRACT

The known interval scale, referred to as the 7.8 scale, has been criticized as an invalid measuring instrument in the form of an attitude scale. It is the purpose of this paper to demonstrate that this scale can produce spuriously inflated correlation coefficients, high reliability, and false significance on statistical tests. The case will be made along two general lines. First, the effects of the scale on reliability, validity, and significance testing will be presented and second the reasoning behind the scale and the method for attaining its values will be discussed. It is concluded that increases in reliability and validity coefficients obtained with the 7.8 scale only confirm that by artificially extending the range of a scale it is possible to increase the correlation coefficient. The application of the Jones and Thurstone scale to the 7.8 scale, whether transcribed correctly or incorrectly, happens to form a non-monotonic transformation, with resultant r and increased chance for significance, which makes the scale appealing on its surface. Reliability and validity, to be useful concepts, must be the reliability and validity of data gathering instruments rather than the reliability and validity of a particular set of scale values. (Author/BJG)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

ED109229

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE NATIONAL INSTITUTE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

TM 004 690 BEST COPY AVAILABLE

A Comment on the "Known-Interval" Scale as a Measure of Attitude

Speech Communication Association

Chicago, Illinois

December 27-30, 1974

Thomas M. Steinfatt

Queens College of CUNY

While its theories, philosophies, research designs, and statistical methods are of great importance to any science, perhaps most fundamental to a science are its measuring instruments. These instruments operationalize the epistemic relationship between the constructs of a science and reality. Without valid measuring devices a science will remain forever in the dark ages. Thus, it is particularly disturbing to find an invalid measuring instrument in the form of an attitude scale beginning to appear in the experimental literature in communication.

Several recent award winning papers at SCA and ICA conventions[1] and at least two articles recently published[2] have used an attitude scale often referred to as a "known interval" scale.[3] While this scale was assuredly inspired by laudable motives, it can produce spuriously inflated correlation coefficients, spuriously high reliability and validity estimates, and spurious significance on statistical tests. It is the purpose of this paper to demonstrate these results, and to argue that the scale should not be used in future studies.

The reasoning will take two general lines. First, the effects of the scale on reliability, validity, and significance testing will be demonstrated. Then the reasoning behind the scale and the method of obtaining its values will be discussed.

Figure 1 about here

The "known interval" scale, referred to hereafter as the 7.8 scale, is reproduced in Figure 1. The parentheses around the numbers indicate that subjects responding to the scale are presented with only the blanks and the

anchoring terms, not with the numerals. The scale was developed because "[it] recognizes the non-equal nature of people's perceptions of attitude scales and is appropriate for use with interval statistics."[4] I assume that the term "interval statistics" refers to parametric tests. Rather than arguing this point in detail, I simply note that while interval scales are helpful if they truly are interval scales, they are not a prerequisite to making a statistical inference based on a parametric test.[5] But ignoring the rationale behind the scale for the moment, let us consider the effects of employing the scale in communication research.

CORRELATION AND TESTS OF SIGNIFICANCE

One of the principal defenses of the 7.8 scale is that it correlates highly with semantic differential scales, higher than does a single seven point scale.

> One item scales have been called notoriously unreliable by many. However, the carefully constructed known-interval scale used in this study had an extremely high correlation with the semantic differential scales which indicates it is not unreliable. Second, it had predictive validity and produced the same findings as the semantic differential-type items. It has two significant advantages: (1) it is much easier to administer than the semantic differential type items, and (2) it yields a lower within error estimate which reduces the likelihood of obscuring significant results when in fact they do exist.[6]

The reliability and predictive validity claimed for the 7.8 scale are due to two sources: (a) the reliability and predictive validity of a "regular" seven point scale (however large that is) and (b) the spurious

increase in these components produced by counting a one unit shift as though the subject had moved 1.8 units. To understand why this is so, consider the effect of extreme scores on the value of a Pearson r. Figure 2 partially illustrates this effect. The scores of the first six subjects

---

Figure 2 about here

---

are negatively correlated, r_6=-.26. When the scores of the seventh subject are included, the correlation increases to +.29. These values are obtained using the ordinary seven point scale. Suppose that in place of 7 we use 7.8. The correlation is increased to +.42. If the 7 is replaced by 10, r_{10}=+.66. By raising the value of this single extreme score, it is possible to increase r as high as one desires. Thus, r_{20}=+.93 and r_{100}=+.997. Changing the values of the first six scores from regular to 7.8 form has a negligible effect on r. Notice that the first six number pairs remain negatively correlated despite changes in the overall r due to the value assigned to the extreme score pair. Notice also that the changes in r are in no way related to any real behavioral event. Subject number seven checked the extreme score for administration X_1 and again for administration X_2. He did this only once. The observed changes in r are due solely to the value assigned to the extreme score after the data have been collected. Had subject number seven checked the seventh blank on X_1 and the first blank on X_2, repeating the above procedure of increasing the scale value of the seventh blank would produce an increasing correlation in the negative direction.

If we interpret the correlations of Figure 2 as test-retest reliability, then it is obvious that we can increase r_{tt} as high as we wish by changing the numerical value assigned to the extreme score. It is also obvious that this increase in r_{tt} has nothing to do with increased reliability. The event

of the experimenter assigning a large number to a scale value is independent of the event of a subject consistently checking the same scale position on repeated measures.

Having demonstrated the effect of changes in extreme score values on r interpreted as reliability, its effect on validity has also been demonstrated. If each of the X_2 values in Figure 2 is multiplied by four, X_2 becomes a scale running from 4 to 28. This is the range of the four summated semantic differential type scales used as a criterion. In the above quotation, the correlation between the seven and the 28 point scales is referred to as reliability. I choose to call it validity. In either case, multiplying the scores by a constant has no effect on r. Thus the arguments in the preceding paragraphs apply directly to $r_{(7)(28)}$ and any increase in $r_{(7.8)(28)}$ over $r_{(7.0)(28)}$ is the spurious result of changing the scale value. It is unrealted to the reliability or validity of the actual data. Skeptics may work these calculations out for themselves.

Predictive validity is claimed for the scale since it "produced the same findings" as did the semantic differential items. This claim is true only to the extent that any single seven interval scaling device measures what four summated scales in semantic differential form measure. To this extent, assigning the numbers "1" to "7" to the data, as in a regular seven interval scale, will have predictive validity. Any predictive validity the 7.8 scale has derives directly from the predictive validity of the general seven point case. But the predictive correspondence of the 7.8 scale to the semantic differential scale will often be less than for the regular seven point scale, since the 7.8 scale can produce spurious significance on statistical tests when such significance does not in fact exist. Consider the data of Figure 3.

Figure 3 about here

Here Y_1 and Y_2 may be interpreted as pretest and posttest scores respectively, of six subjects on a seven point scale. The t value for this data is 2.318 with 5 degrees of freedom which is not significant at the .05 level, two-tailed. If these same data are rescored as advocated by the 7.8 scale, they become significant ($t=2.95$, $df=5$, $p < .05$, two-tailed). This is not an isolated instance. It can occur on any statistical test (t, Scheffé, Anova, etc.) with any K sets of data, provided only that one set of data have relatively few scores of "seven" and that the other set(s) of data has (have) many, and that the mean of the "non-seven" scores in one set of data is fairly close to the mean of the "non-seven" scores in the other set(s). Such conditions are frequently met in empirical data. The 7.8 scale is thus capable of producing spuriously significant results.

THE APPLICATION OF JONES AND THURSTONE TO ATTITUDE MEASUREMENT

Given the effects of the 7.8 scale as demonstrated above, the reasoning behind this scale can be examined. Two questions are asked in this section. First, are the findings of Jones and Thurstone [7] applicable to attitude measurement? Second, were Jones and Thurstone's findings applied correctly in the case of the 7.8 scale?

Is Jones and Thurstone Applicable in General?

To answer the first question, consider what Jones and Thurstone did. During the early 1950's they administered questionnaires containing 51 descriptive words and phrases (see Table 1) to 905 enlisted personnel at Fort Lee, Virginia, who were asked to give their meanings for these words

and phrases in terms of the amount of like or dislike the words and phrases indicate in preference for foods, such as creamed corn. These indicated preferences were marked on a nine point scale which was anchored at both ends and in the middle with the phrases "Greatest Dislike," "Neither Like Nor Dislike," and "Greatest Like." The symbols "-4," "-3," ... "+2," "+3," "+4" appeared above the nine blanks on each scale. The responses were scaled by the method of successive intervals which produces normal deviates (z scores) for each item. These normal deviates may then be interpreted as a continuum of meaning for the 51 items, as in Table 1 which is taken, in part, from Table 2 in Jones and Thurstone.[8]

Burgoon[9] selected the seven underlined words in Table 1 as anchors for the 7.8 scale. The values in the scale result from adding a constant (4.1) to the z scores (scale values) obtained by Jones and Thurstone.

There is a general problem of order effects associated with the results of Jones and Thurstone. All 905 subjects responded to all 51 items in exactly the same order. Thus the effects of practice and fatigue are inherently confounded with the rating of each item. The lack of counter balancing for order could, in my opinion, result in a rejection of this paper should it be submitted for publication today. Thus, I believe that the results of Jones and Thurstone are of dubious value for attitude scaling purposes.

Beyond this general criticism, Jones and Thurstone comment on the applicability of their results. They state that their results "might be generalized to the extent that the phrases useful for defining successive intervals on a food preference schedule might also be useful for defining intervals on schedules assessing preferences for other consumer goods."[10] Note that the application of these results to the 7.8 scale has not been in

the area of consumer goods.

Was Jones and Thurstone Applied Correctly in This Case?

Whether or not the reader accepts the general applicability of Jones and Thurstone to attitude scaling, the application in the particular case of the 7.8 scale is incorrect. In particular, the application of Jones and Thurstone to the 7.8 scale was done in reverse (i.e., backwards). Instead of choosing anchors for a regular seven point scale based on the Jones and Thurstone results, the seven scale values were assigned on the basis of the chosen anchors. That is, instead of selecting equally spaced words and assigning them as anchors to seven equally spaced numerals, seven unequally spaced words were selected and the scale values changed to conform with these unequal intervals. Since the fallacy in this logic may not be immediately apparent, let us form another attitude scale called the TIC scale (Tongue in Cheek) using the same logic. For my TIC scale I will choose the same six anchors used by the 7.8 scale for the first six scale positions (Terrible, Bad, etc.). But in place of "Excellent" for the seventh scale value, I choose "Best of all" for the TIC scale, resulting in a seven point scale from 1.0 to 10.25 (See Figure 4). If the 7.8 scale is good in terms

---

Figure 4 about here

---

of producing reliability, validity, and significance, the TIC scale is spectacular. In terms of reliability and validity, its results may be compared with the 7.8 scale from the calculations presented in the first section of this paper (r_{10}=.66, $r_{7.8}$=.42). TIC will also out-perform the 7.8 scale in getting significance out of a given set of data.

Unfortunately, even if this application of Jones and Thurstone to the 7.8 scale and the TIC scale were legitimate, which it is not, the data would

still not be interpretable. A crucial assumption of both 7.8 and TIC is that a change from "Good" to "Excellent," or to "Best of all," corresponds to a subject's perception of a change of 1.8 (7.8-6.0), or 4.25 (10.25-6.0), units on his perceptual scale, compared with approximately one unit change between each of the lower six scale values. But there is no way of knowing what part of the scale stimulus the subject is responding to. Is he/she responding to the equally spaced intervals as they appear on the page, or to the anchors below the blanks? The TIC and 7.8 scales assume that people are responding to the anchors rather than to the equal spacing on the scale. The scale gives subjects an ambiguous choice. If they really do perceive the differences between words as they are scaled, then these distances conflict with the equal spaces between the words on the paper. Which is the subject to choose? Which does each subject choose? Since there is no way to know this, the 7.8 scale is ambiguous and necessarily produces ambiguous (and thus uninterpretable) results.

Finally, there are a number of errors in the transformation of the data between Jones and Thurstone and the 7.8 scale. In the reproduced Table from Jones and Thurstone[11] there are ten copying errors. These errors are presented in Table 1. One of these errors occurs on the word "Poor" which

---

Table 1 about here

---

is scaled at -1.55 by Jones and Thurstone, but appears as -1.35 in Burgoon.[12] The effect of this error and an addition error is illustrated in Table 2. When 4.1 is added to 0.02 the result should be 4.12 which rounds to 4.1,

---

Table 2 about here

---

not 4.0. If 4.1 is added to -1.55, the result is 2.55 which rounds to 2.6, not 2.9. There is an addition error here as well as a copying error, since the copying error accounts for only 2 of the .3 discrepancy. The differences between the table values reported by Burgoon and those listed by Jones and Thurstone also affect the language intensity manipulations, such as found in Burgoon and Chase, and Burgoon.[13] For example, Jones and Thurstone did not test the phrase "Mighty favorable" which is employed in the language intensity manipulation with a value of 2.81. This score was achieved by the phrase "Highly favorable."

CONCLUSION

This paper has attempted to show that increases in reliability and validity coefficients obtained with the 7.8 scale have nothing to do with observed data and, thus, have nothing to say about observed data. The increase only confirms that by artificially extending the range of a scale it is possible to increase a correlation coefficient. A monotonic transformation which preserves the relative intervality of the data would not affect r. It is because the Jones and Thurstone scale values, whether transcribed correctly or incorrectly, happen to form a non-monotonic transformation, with resultant increased r and increased chance for significance, which makes the scale appealing on its surface.

Rather than trying to adjust the values on the instruments after the horse is out of the barn (after the data have been collected) researchers in communication should develop more valid and reliable instruments for measuring important communication variables. This was assuredly the intent behind the formation of the 7.8 scale. But reliability and validity to be useful concepts must be the reliability and validity of data gathering instruments, not the reliability and validity of a particular set of scale values.

Footnotes

1. Michael Burgoon and Lyle B. King, "The Mediation of Resistance to Persuasion Strategies by Language Variables and Active-Passive Participation," paper presented on the "Top Three Plus Critics" program of the Interpersonal and Small Group Interaction Division at the Speech Communication Association Annual Meeting, New York, November 1973; Judee K. Heston, "Conflicting Information, Attitude and Message Variables as Predictors of Learning and Persuasion," paper presented on the "Top Three Plus Critics" program of the Interpersonal Communication Division at the Annual Meeting of the International Communication Association, New Orleans, April 1974.

2. Michael Burgoon and Gerald R. Miller, "Prior Attitude and Language Intensity as Predictors of Message Style and Attitude Change Following Counterattitudinal Advocacy," Journal of Personality and Social Psychology, 20 (1971), 246-253; Michael Burgoon and Lyle B. King, "The Mediation of Resistance to Persuasion Strategies by Language Variables and Active-Passive Participation," Human Communication Research, 1 (1974), in press.

3. This scale first appeared in Michael Burgoon, "Prior attitude and language intensity as predictors of message style and attitude change following counterattitudinal communication behavior," Unpublished Doctoral Dissertation, Michigan State University, 1970.

4. Burgoon and King, 1973,8.

5. Frederic M. Lord, "On the Statistical Treatment of Football Numbers," American Psychologist, 8(1953), 750-751; C. Alan Boneau, "A Note on Measurement Scales and Statistical Tests," American Psychologist, 16 (1961), 160-261; John Gaito, "Scale Classification and Statistics," Psychological Review, 67(1960), 277-278; Norman H. Anderson, "Scales and Statistics: Parametric and Nonparametric," Psychological Bulletin, 58(1961), 305-316.

6. Burgoon and King, 1973, 13.

7. Lyle V. Jones and L. L. Thurstone, "The Psychophysics of Semantics: An Experimental Investigation," Journal of Applied Psychology, 39 (1955), 31-36. The values for the scale points in the 7.8 scale were calculated from the results of this study.

8. Jones and Thurstone, 33.

9. Burgoon, 1970, 21.

10. Jones and Thurstone, 31.

11. Burgoon, 1970, 21.

12. Ibid.

13. Michael Burgoon and Lawrence J. Chase, "The Effects of Differential Linguistic Patterns in Messages Attempting to Induce Resistance to Persuasion," Speech Monographs, 40(1973), 1-7; Burgoon, 1970.

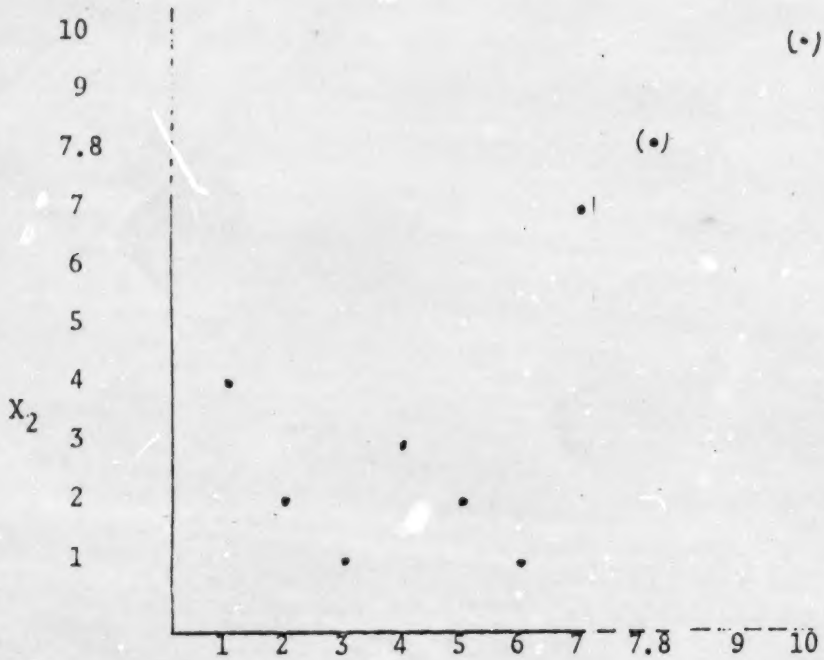
"Regular" Seven Point Scale

Terrible	Bad	Poor	Neutral	Fair	Good	Excellent
(1.0)	(2.0)	(3.0)	(4.0)	(5.0)	(6.0)	(7.0)

"Known Interval" or "7.8" Scale

Terrible	Bad	Poor	Neutral	Fair	Good	Excellent
(1.0)	(2.1)	(2.9)	(4.0)	(4.9)	(6.0)	(7.8)

Figure 1. Comparison of the usual seven point scale values with values proposed for the 7.8 scale.



X_1	1	2	3	4	5	6	7
X_2	4	2	1	3	2	1	7

$r_6 = -.26$

7	$r_7 = +.29$
7.8	$r_{7.8} = +.42$
10	$r_{10} = +.66$
20	$r_{20} = +.93$
100	$r_{100} = +.997$

Figure 2. Illustration of the effect of extreme scores on a Pearson r.

Y_1	3	2	1	1	4	2
Y_2	7	4	7	7	4	1

$$t = \frac{2.833}{1.222} = 2.318$$

N.S.D.
Two-Tailed

Regular seven point scale

Y_1'	2.9	2.1	1	1	4	2.1
Y_2'	7.8	4	7.8	7.8	4	1

$$t = \frac{3.583}{1.215} = 2.949$$

$p < .05$
Two-Tailed

Same data using 7.8 scale

Figure 3. Illustration of spurious significance produced on t-test for two related samples using the 7.8 scale.

Terrible	Bad	Poor	Neutral	Fair	Good	Best of all
(1.0)	(2.1)	(2.9)	(4.0)	(4.9)	(6.0)	(10.25)

Figure 4. The TIC attitude scale.

Table 1

Order of presentation, name of item, and scale value obtained by Jones and Thurstone (1955, Table 2, 33) compared with reproduction of this table in Burgoon (1970, Table 1, 21)

Jones and Thurstone Item No.	Jones and Thurstone Item	J & T Scale Value	Scale Value in Burgoon
25	Best of all	6.15	Same
41	Favorite	4.68	"
23	Like extremely	4.16	"
1	Like intensely	4.05	"
30	Excellent	3.71	"
12	Wonderful	3.51	3.31
50	Strongly like	2.96	Same
17	Like very much	2.91	2.90
42	Mighty fine	2.88	Same
14	Especially good	2.86	"
13	Highly favorable*	2.81	"
20	Like very well	2.60	"
21	Very good	2.56	2.36
26	Like quite a bit	2.32	2.31
19	Enjoy	2.21	Same
8	Preferred	1.98	1.96
35	Good	1.91	Same
6	Welcome	1.77	"
48	Tasty	1.76	"
27	Pleasing	1.58	"
40	Like fairly well	1.51	"
24	Like	1.35	1.38
29	Like moderately	1.12	Same
45	OK	.87	"
31	Average	.86	"
5	Mildly like	.85	"
37	Fair	.78	"
33	Acceptable	.73	"
51	Only fair	.71	"
15	Like slightly	.69	"
11	Neutral	.02	Same
10	Like not so well	-.30	"
28	Like not so much	-.41	"
2	Dislike slightly	-.59	"
22	Mildly dislike	-.74	"
7	Not pleasing	-.83	"
47	Don't care for it	-1.10	"
39	Dislike moderately	-1.20	"
36	Poor	-1.55	-1.35
46	Dislike	-1.58	Same
44	Don't like	-1.81	"
49	Bad	-2.02	"
18	Highly unfavorable	-2.16	"
3	Strongly dislike	-2.37	"
9	Dislike very much	-2.49	"
34	Very bad	-2.53	"
43	Terrible	-3.09	"
16	Dislike intensely	-3.33	-3.31
32	Loath	-3.76	Same
38	Dislike extremely	-4.12	-6.22
4	Despise	-6.44	Same

*In the Burgoon (1970) Table, this item is listed as "Mighty favorable."

Table 2

Effect of a copying error and incorrect addition on two of seven values in the 7.8 scale.

	Terrible	Bad	Poor	Neutral	Fair	Good	Excellent
Jones and Thurstone Scale Values	-3.09	-2.02	-1.55	0.02	0.78	1.91	3.71
7.8 Scale Values	1.0	2.1	2.9	4.0	4.9	6.0	7.8
Jones and Thurstone Scale Values +4.1	1.01	2.08	2.55	4.12	4.88	6.01	7.81
J&T + 4.1 Rounded	1.0	2.1	2.6	4.1	4.9	6.0	7.8